NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AN EMPIRICAL EVALUATION OF A FACTOR EFFECTS SCREENING PROCEDURE FOR EXPLORING COMPLEX SIMULATION MODELS

by

Kerry N. Bosché

June 2006

Thesis Advisor:	Susan M. Sanchez
Second Reader:	Hong Wan

Approved for public release; distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June/2006	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE An Empirical Evaluation of a Factor Effects Screening Procedure for Exploring Complex Simulation Models		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Kerry N. Bosché			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

Screening experiments are procedures designed to identify the most important factors in simulation models. Previously proposed one-stage procedures such as sequential bifurcation (SB) and controlled sequential bifurcation (CSB) require factor effects to be arranged according to estimated sign or magnitude prior to screening. FF-CSB is a two-stage screening procedure for simulation experiments proposed by Sanchez *et al.* (2005) which uses an efficient fractional factorial experiment to estimate factor effects automatically, removing the need for pre-estimation. Empirical results show that FF-CSB classifies factor effects as well as CSB in fewer runs when factors are only grouped by their sign (positive or negative). In theory, the procedure can achieve more efficient run times when factors are also sorted by estimated effect after the first stage. This analysis tests the efficiency and performance characteristics of a *sorted* FF-CSB procedure under a variety of conditions and finds that the procedure classifies factors as well as *unsorted* FF-CSB with significant improvement in run times. Additionally, various model- and user-determined scenarios are tested in an initial attempt to parameterize run times against parameters known or controlled by the modeler. Further experimentation is also suggested.

14. SUBJECT TERMS Factor effects screening, screening experiments, controlled sequential bifurcation, simulation experiments			**15. NUMBER OF PAGES** 53
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited.

AN EMPIRICAL EVALUATION OF A FACTOR EFFECTS SCREENING PROCEDURE FOR EXPLORING COMPLEX SIMULATION MODELS

Kerry N. Bosché
Ensign, United States Navy
B.S., United States Naval Academy, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN APPLIED SCIENCE (OPERATIONS RESEARCH)

from the

NAVAL POSTGRADUATE SCHOOL
June 2006

Author: Kerry N. Bosché

Approved by: Susan M. Sanchez
Thesis Advisor

Hong Wan
Second Reader

James N. Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Screening experiments are procedures designed to identify the most important factors in simulation models. Previously proposed one-stage procedures such as sequential bifurcation (SB) and controlled sequential bifurcation (CSB) require factor effects to be arranged according to estimated sign or magnitude prior to screening. FF-CSB is a two-stage screening procedure for simulation experiments proposed by Sanchez *et al.* (2005) which uses an efficient fractional factorial experiment to estimate factor effects automatically, removing the need for pre-estimation. Empirical results show that FF-CSB classifies factor effects as well as CSB in fewer runs when factors are only grouped by their sign (positive or negative). In theory, the procedure can achieve more efficient run times when factors are also sorted by estimated effect after the first stage. This analysis tests the efficiency and performance characteristics of a *sorted* FF-CSB procedure under a variety of conditions and finds that the procedure classifies factors as well as *unsorted* FF-CSB with significant improvement in run times. Additionally, various model- and user-determined scenarios are tested in an initial attempt to parameterize run times against parameters known or controlled by the modeler. Further experimentation is also suggested.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

Susan M. Sanchez, Department of Operations Research, Naval Postgraduate School for proposing this study and providing guidance throughout.

Assistant Professor Hong Wan, Department of Industrial Engineering, Purdue University.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Screening experiments are procedures designed to identify the most important factors in simulation models. To use computing resources efficiently, an analyst can then focus resources on higher resolution study of these factors. Group screening methods are a family of screening procedures that test the effect of groups of factors. Depending on the group effect, the group is either classified unimportant or split and added to a queue for further estimation until all factors are classified.

Previously proposed screening procedures include sequential bifurcation (SB) and controlled sequential bifurcation (CSB). CSB uses sequential sampling to guarantee a user-specified maximum Type I error for factor effects below a threshold Δ_0 and a user-specified minimum power of detection for factor effects above a threshold Δ_1. Both thresholds are also user-specified. However, both SB and CSB require the signs of factor effects to be known prior to screening, and perform best if the factors are initially arranged according to the magnitudes of their effects. By splitting the factors into a positive group and a negative group, the analyst can avoid misclassification of potentially important factors of opposite sign. Unfortunately, estimation of factor effects has shown to be unreliable and time-consuming in practice. To address this, Sanchez et al. (2005) propose FF-CSB, a two-phase screening procedure for simulation experiments which uses an efficient fractional factorial experiment to estimate factor effects automatically, removing the need for pre-determining the factor effect signs or orderings. Empirical results show that FF-CSB classifies factor effects as well as CSB in fewer runs when factors are grouped only by their sign (positive or negative). In theory, the procedure can achieve shorter run times when factors are also sorted by estimated effect after the first stage.

This analysis tests the efficiency and performance characteristics of a *sorted* FF-CSB procedure under a variety of conditions. Preliminary results show that the procedure classifies factors as well as *unsorted* FF-CSB with significant improvement in run times (up to 26% improvement in this analysis). Additionally, *s*FF-CSB becomes more efficient relative to *u*FF-CSB as more factors are added. The effect on procedure performance of

other arrangements of factor effects is also examined. The results show that while correct detection rates are guaranteed by CSB, run times are dependent on several factors.

Results show that runtimes increase when threshold values are placed closer together. For the analyst, this represents a tradeoff between run times and better overall classification rates of factor effects. In situations of sparsity, where factor effects are either zero or the experimental maximum, run times are considerably smaller than in the non-sparse case. The number of negative factors relative to the total number of factors does not appear to have an effect on run times.

Since response variance is a major consideration for sequential methods such as CSB and FF-CSB, the effect of increased variance is tested. Preliminary results suggest that run times grow with the variance, so that a model response with twice the variance of a base case will take approximately twice as many runs to complete FF-CSB.

The effect of variance is also evaluated by parameterizing response variance as a function of the mean response. Response standard deviation is set as proportional to the response value times a multiplier m. This represents an extremely difficult scenario for the CSB phase, which utilizes hypothesis testing, since every group effect, regardless of magnitude, has an equal probability of failing to reject when compared to zero. The results show that run times are considerably larger in these situations. Additionally, run times are larger in the non-constant case when compared to a case where variance is constant and mean variances for the two cases are roughly equivalent.

For the analyst, these results not only present an improved screening procedure, but give a more complete description of performance under certain conditions. Further research is suggested that will provide insight into the behavior of screening procedures in an effort to provide a practical and well-understood screening tool to analysts.

I. INTRODUCTION

Factor effect screening procedures are a class of algorithms that can be applied to study the behavior of complex systems in order to quickly find the factors that have the largest overall effects on performance. These procedures are valuable for defense analysts because they provide systematic ways of investigating simulation models of military operations. Screening of factor effects is especially useful when the number of factors is large. When a long list of potentially important factors can be quickly trimmed to a short list of demonstrably important factors, this allows analysts to apply a greater portion of their time, effort, and computing resources toward higher-resolution experiments that focus on the factors that matter. Efficient and reliable screening procedures applied to simulation experiments can greatly improve the efficiency of the modeling and decision-making process, but even so they can require substantial computing resources, especially when the number of factors is high. It is important to the analyst that the screening procedure produces reliable and robust results while using computer memory, clock time, or a given number of model runs efficiently.

Several types of screening procedures exist, including fractional factorials and group screening methods such as sequential bifurcation (SB). Two-level fractional factorial designs estimate the metamodel by sampling a relatively small number of input combinations (design points) at the extremes of the modeling space. The analyst assumes that only main effects are important, or that factors with strong interactions can be identified via their main effects. Group screening methods in simulation are analogous to physical group screening methods such as drug or disease testing methods, where portions of individual samples in a batch can be combined and tested at once, and then further subdivided and tested in the event that the group tests positive. This event is statistically unlikely if all the elements of the group are, indeed, negative. As applied to simulation experiments, group screening algorithms aggregate factors into groups and use an efficient experimental design to test the effect of the entire group. Groups that do not pass procedure-specific criteria for having a significant effect on outcomes are eliminated from consideration, while groups that have a significant effect are further subdivided and tested until individual factors remain. Group screening procedures generally eliminate

unimportant factors relatively quickly and progress sequentially to a smaller number of increasingly important factors.

Using SB as proposed by Bettonvil and Kleijnen (1997), important factors can be identified in a relatively small number of runs assuming that they are sparse (relatively few factors are important and factors are either important or unimportant, not marginally one or the other) and that the response surface can be reasonably approximated by a non-stochastic main-effects model such as

$$Y = \beta_0 + \sum_{i=1}^{K} \beta_i x_i \,. \tag{1}$$

Here, Y represents the simulation model output, K represents the number of factors under investigation, factor effects are represented by β_i (i=1,…,K), and x_i denotes the level of the ith factor (i.e., the value for the ith simulation input). A further assumption of SB is that the signs of the effects are known before experimentation begins, meaning that any factors associated with negative effects can be recoded so the signs are all assumed to be non-negative. While SB can be much more efficient than fractional factorial designs, the assumptions are not representative of most useful, comprehensive models. Even in non-stochastic models, a main-effects model may be insufficient to capture significant non-linearity.

Cheng (1997) expands SB to stochastic response models by assuming that errors are identically, normally distributed. A stochastic main effects model is assumed.

$$Y = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + \varepsilon, \text{ where } \varepsilon \sim \mathrm{N}\left(0, \sigma^2\right) \tag{2}$$

As in equation (1), Y represents the simulation model output, and K represents the number of factors under investigation. Factor effects are represented by β_i, and x_i denotes the level of the ith factor. Group effects are classified using an indifference zone approach in an attempt to control power and probability of error. The analyst is required to supply a threshold value Δ, and factor effects estimated to be below Δ during the procedure are considered unimportant. The author notes that although the stochastic nature of the response is addressed, there is no easy way to determine whether all factors are correctly classified.

Other screening procedures have been proposed specifically for stochastic simulation studies. Like SB, though, none of these approaches give theoretical guarantees of the correctness of factor classifications or the number of runs required, and they typically assume equal variance across different factor settings, which is rare for simulations of complex systems. Although screening procedures in general are orders of magnitude times more efficient than full factorial experiments, and allow the analyst to guide further detailed analysis rather than arbitrarily simplifying the model, the lack of guarantees on classification of factors leaves more to be desired.

In response, Wan, et al. (2003) propose controlled sequential bifurcation (CSB), an adaptation of SB which guarantees the probability of correct factor classification. In CSB, the analyst supplies two thresholds. The lower threshold (Δ_0) is the lower bound on the true (but unknown) factor effect that the decision-maker considers to be *important*. Factors with effects below Δ_0 are considered *unimportant.* The higher threshold (Δ_1) is the equivalent bound for factors to be considered *critical*. CSB uses the same group testing and subdividing procedure as SB, except that hypothesis testing is used at every step to guarantee a maximum probability that an unimportant factor is classified as important (type I error) and a minimum probability that critical factors are classified as important (power). Factors belonging to groups whose group effect is sufficiently small are classified as *unimportant* and do not require further testing. If a group containing two or more factors has a large group effect, then it is split into smaller groups for further testing. If a lone factor has a sufficient effect output, it is classified as *important*. When all factors have been classified, the procedure is finished.

Like SB, CSB assumes that the sign of factor effects are previously known or estimated in order to recode negative factors as positive, avoiding cross-cancellation and elimination of critical factors. In practice, however, the assumption that the signs can be correctly classified on a consistent basis, even by a subject-matter expert, is optimistic, especially in the presence of many hundreds or thousands of factors and secondary or tertiary effects. Lucas et al. (2002) show that even experts may be unable to pre-identify the few most important factors and that factor effect magnitudes often do not behave as predicted. The likelihood that subject-matter experts can correctly pre-order the factors

according to their effects is much smaller, making it extremely unlikely that the potential accuracy and efficiency of CSB can be realized.

To eliminate the uncertainties associated with pre-identification of factor effects, Sanchez, Wan, and Lucas (2005) propose a combination of a fractional factorial experiment with the CSB procedure described above into a two-stage screening procedure (FF-CSB). The initial fractional factorial step estimates the signs and magnitudes of the effects not as a decision-making tool in itself, but in order to create the first subdivision in CSB. Previously-determined qualitative guesses of the signs and relative orderings of the effect magnitudes are no longer needed since numerical estimates based on the model at hand are provided. Using a stochastic main-effects model to simulate the response of a notional model being screened, the authors show that even when the factors are grouped only by sign, the chance of cross-cancellation and disposal of important or critical factors are substantially reduced as compared to CSB. In the specific case where half of the factor effects are negative, CSB as tested is unable to identify *any* factors as important or critical, whereas FF-CSB identifies at least 99.8% of *critical* and *unimportant* factors correctly. Despite the extra fractional factorial sampling during the first phase of the procedure, FF-CSB is extremely efficient. The authors also perform preliminary analysis on the relative performance of FF-CSB when the factors are not only grouped according to sign but sorted according to magnitude after the initial step in order to increase efficiency. Empirical testing has shown that both stochastic and non-stochastic SB screen factors in fewer runs when factors are sorted in order according to estimated magnitude either by a subject-matter expert or by previous screening. This arrangement, intuitively, saves simulation runs. Unimportant factors will be grouped together toward one end of the group, and a substantial portion will be eliminated before the procedure must handle more than one group containing an important factor. Sanchez, Wan, and Lucas (2005) show that, over a varying number of factors, the sorted FF-CSB requires *at most* 79% of the runs required by the unsorted version and that relative efficiency of the sorted procedure increases with the number of factors.

This paper expands the preliminary sorting experiment executed by Sanchez, Wan, and Lucas (2005) and uses the same model response function (described in Section 3) to explore the effect of sparsity, relative threshold values, prevalence of negative

effects, and heterogeneity of error variance on model performance. The relationship between factor effect magnitude and classification of important factors is also investigated. The results show that given an identical group of factors, a sorted FF-CSB procedure correctly classifies factors at least as well as the unsorted procedure with a 10% to 30% improvement in run time. The results also show that while sparsity, closer threshold values, and heterogeneous errors can have a large detrimental effect on run times, CSB error and power specifications guarantee relatively constant probabilities of detection for a given magnitude when threshold values are identical. For a given set of factors, though, the improvement in run time over the unsorted procedure verifies the advantage of the extra sorting step. The significance to modelers of each of these considerations is discussed, and further research is suggested.

THIS PAGE INTENTIONALLY LEFT BLANK

II. FF-CSB PROCEDURE

In this analysis, both the CSB procedure of Wan et al. (2003, 2006) and the FF-CSB procedure described by Sanchez et al. (2005) assume a stochastic main-effects metamodel for simulation response. The model contains a user defined number of factors (*K)* with coefficients $\beta_1,...,\beta_k$ and is defined below:

$$Y = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + \varepsilon . \tag{3}$$

Y represents the simulation model output, and errors are distributed $\varepsilon \sim N(0, \sigma_{\mathbf{x}}^2)$. Note that the variance of *Y* can depend on the value of $\mathbf{x} = (x_1,...,x_k)$. Although, in practice, the assumption of a main effects model in simulation is not likely to hold over the entire factor space, it is more likely to be a reasonable assumption in a small region of the factor space, where the factor levels are limited. In this analysis, factor effects are estimated only at their extreme points. The shape of the response curve between these points is assumed to be either strictly linear or dominated by linear main effects, even if interactions are present. A more precise picture of the response is unnecessary, since the purpose of a screening experiment is to identify factors that, in general, have the most effect on outcomes and are therefore good candidates for higher resolution study.

Also, since relatively few factors are assumed to have a significant (*important*) effect on the simulation output, the addition of unimportant factors can make the factor space much larger while the response does not depart considerably from that of the original metamodel.

The FF-CSB procedure used in this analysis and displayed in Table 1 is that of Sanchez et al. (2005). To initialize groups of factors to be tested, two LIFO queues, NEG and POS, are created to hold factors according to the sign of their effect as estimated in the fractional factorial step: one each for factors with positive and negative effects. The fractional factorial experiment is run, and the factors are ordered according to their estimated magnitude $(\hat{\beta}_i)$ such that

$$\hat{\beta}_{[1]} \le ... \le \hat{\beta}_{[z]} < 0 \le \hat{\beta}_{[z+1]} \le ... \le \hat{\beta}_{[K]} . \tag{4}$$

Factors with negative estimated effects ($[1],...,[z]$) are placed in the NEG queue and factors with non-negative estimated effects ($[z+1],...,[k]$) are placed in the POS queue. This completes Phase 1.

Table 1. FF-CSB Procedure

Initialization:
Create two empty LIFO queues for groups, NEG and POS.
Phase 1:
Conduct a saturated or nearly-saturated fractional factorial experiment and estimate $\hat{\beta}_1,...,\hat{\beta}_k$. Order the estimates so that $\hat{\beta}_{[1]} \le ... \le \hat{\beta}_{[z]} < 0 < \hat{\beta}_{[z+1]} \le ... \le \hat{\beta}_{[K]}$. Add factors {[1], . . . , [z]} to the NEG queue, and factors {[z + 1], . . . , [K]} to the POS queue.
Phase 2:
For queue = POS and queue = NEG, do
While queue is not empty, do
Remove: Remove a group from the queue.
Test:
Unimportant:
If the group effect is unimportant ($<\Delta_0$), then classify all factors in the group as *unimportant*.
Important (size=1):
If the group effect is important ($>\Delta_0$) and of size 1, then classify the factor as *important*.
Important (size>1):
If the group effect is important ($>\Delta_0$) and the size is greater than 1, then split the group into two subgroups such that all factors in the first subgroup have smaller [i]'s (ordered indices) than those in the second subgroup. Add each subgroup to the LIFO queue.
End Test
End While
End For

Phase 2 applies CSB to each queue in turn. For each replication of the CSB procedure, a group of factors is removed from the queue and the combined effect of the entire group is tested. If the group effect is *unimportant*, or below the Δ_0 threshold, every factor in the group is labeled as unimportant. Groups whose effects are *important* are handled according to size. If the group contains only one factor, the factor is labeled as *important*, according to the magnitude of the estimated effect. If multiple factors make up the group, it is further subdivided into two groups such that the estimated magnitudes are all smaller in one group than in the other, preserving the initial ordering by magnitude. Each new group is added back to the queue.

At each step, when a group of factors is tested, CSB uses hypothesis testing based on a sample of model outputs (Y) at a given design point to place statistical guarantees on factor classification. If all factors are correctly classified into NEG and POS groups after

Phase 1, then for any factor with $|\beta_i| < \Delta_0$, CSB guarantees that the probability of classification as *important* (type I error) is less than a user specified α. Similarly, for any factor with $|\beta_i| > \Delta_1$, the probability of classification as *important* is greater than a user controlled γ. This relationship is displayed graphically in Figure 1. Ideally, observed detection rates should rise with effect magnitude. Effects in the *unimportant* range are identified as *important* at a rate no greater than α. Similarly, effects in the *critical* range are identified as *important* at a rate no less than γ. The default sample size (n_0) is initially small, usually no more than five, and additional samples are taken only as needed to save runs. Wan et al. (2003, 2006) also test a version of CSB where there is no initial minimum sample size and runs are taken one at a time until the effect can be classified. Their results show that this procedure is, on average, more efficient, requiring fewer runs in most cases.

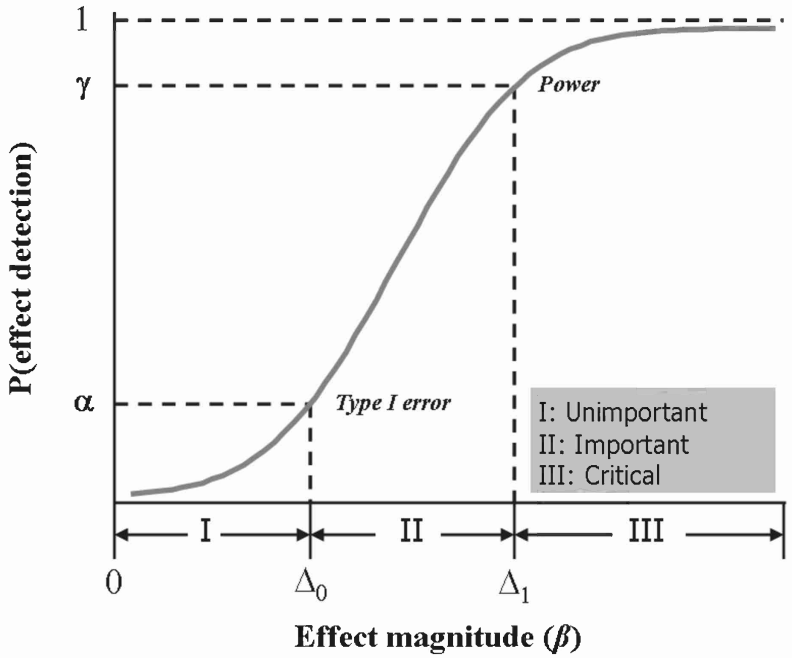

Figure 1. CSB Guarantees the Probabilities That Critical and Unimportant Factors are Correctly Classified (from Wan et al. 2006).

In their initial work on FF-CSB, Sanchez et al. (2005) explore the performance of FF-CSB under the assumption that only the *signs* of the factor effects can be estimated. This is an attempt to establish whether significant performance and efficiency improvements can be achieved only by reducing the probability of cross-

cancellation. The result is that factors are separated only by the estimated signs, not also ordered by magnitude. The authors find that the performance of CSB, summarized in the 'CSB' column of Table 2, is highly dependent on the proportion of negative effects in the experiment. Although the procedure exceeds power and error specifications when none or a small number of effects are negative, the correct classification rates across all categories deteriorate as the number of negative factors approaches the number of non-negative factors. When half of the factors are negative, the procedure is unable to produce useful results. The number of runs required for CSB also decreases as more negative factors are introduced because more factors are discarded, often incorrectly, due to cancellation.

In general, while CSB performs well when the conditions are optimal, it cannot classify factors reliably in the general sense, and any gains in run requirements as negative factors increase are at the cost of accuracy.

In contrast, due to separation by estimated sign after the first stage, the *u*FF-CSB procedure is not as susceptible to cancellation between positive and negative factors during estimation. The procedure, whose performance is summarized in the '*Unsorted* FF-CSB' column of Table 2, exceeds power and error specifications in all cases, regardless of the number of negative effects. *Critical* and *unimportant* effects are correctly classified at least 99.7% of the time, and correct classification rates for *important* factors range from 42.7% to 80.9%. However, the number of runs required for *u*FF-CSB is generally larger. This is generally true for one primary reason: since the *u*FF-CSB has grouped factors by their estimated signs, far fewer factors are discarded as *unimportant* due to cancellation. Consequently, fewer factors are classified per bifurcation step and more steps are required.

Table 2. Performance of CSB and *u*FF-CSB (from Sanchez et al., 2005)

Pattern of β values	K	CSB Correct Class. Prop. Crit.	Imp.	Unimp.	Avg. Runs	*Unsorted* FF-CSB Correct Class. Prop. Crit.	Imp.	Unimp.	Avg. Runs
None Negative	7	1.000	0.788	0.999	100	1.000	0.809	0.999	110
	15	0.999	0.425	1.000	248	1.000	0.432	1.000	268
	31	1.000	0.503	1.000	610	0.998	0.493	0.999	656
	63	1.000	0.492	1.000	1,488	0.999	0.490	1.000	1,563
	127	1.000	0.506	1.000	3,559	1.000	0.507	1.000	3,692
	255	1.000	0.495	1.000	8,192	1.000	0.497	1.000	8,704
	511	1.000	0.500	1.000	19,099	1.000	0.497	1.000	19,528
Small Negative	7	1.000	0.788	0.999	100	1.000	0.809	0.999	110
	15	0.998	0.401	1.000	241	0.999	0.431	1.000	251
	31	0.991	0.458	1.000	581	0.999	0.498	0.999	605
	63	0.994	0.453	1.000	1,424	1.000	0.487	1.000	1,461
	127	0.993	0.469	1.000	3,428	1.000	0.506	1.000	3,421
	255	0.992	0.460	1.000	7,824	1.000	0.496	1.000	8,024
	511	0.992	0.461	1.000	17,958	1.000	0.499	1.000	18,132
Medium Negative	7	0.972	0.671	1.000	92	1.000	0.804	0.999	100
	15	0.909	0.332	1.000	202	0.999	0.432	1.000	250
	31	0.879	0.384	1.000	491	0.999	0.502	0.999	586
	63	0.880	0.381	1.000	1,169	0.999	0.494	1.000	1,407
	127	0.879	0.393	1.000	2,811	1.000	0.508	1.000	3,307
	255	0.878	0.382	1.000	6,568	1.000	0.496	1.000	7,550
	511	0.877	0.386	1.000	15,361	1.000	0.499	1.000	17,703
Large Negative	7	0.744	0.297	1.000	66	1.000	0.785	0.999	100
	15	0.684	0.142	1.000	140	0.999	0.427	1.000	234
	31	0.664	0.164	0.999	339	0.997	0.495	0.999	558
	63	0.663	0.187	1.000	781	0.999	0.489	1.000	1,329
	127	0.662	0.201	1.000	1,827	1.000	0.505	1.000	3,171
	255	0.636	0.185	1.000	4,219	1.000	0.494	1.000	7,440
	511	0.661	0.200	1.000	10,041	1.000	0.499	1.000	17,595
Half Negative	7	0.000	0.000	1.000	10.7	1.000	0.770	1.000	99
	15	0.000	0.000	1.000	10.7	0.999	0.431	1.000	231
	31	0.000	0.000	1.000	10.6	0.998	0.494	0.999	556
	63	0.000	0.000	1.000	10.6	0.999	0.489	1.000	1,313
	127	0.000	0.000	1.000	10.7	1.000	0.507	1.000	3,143
	255	0.000	0.000	1.000	10.7	1.000	0.495	1.000	7,441
	511	0.000	0.000	1.000	10.6	1.000	0.499	1.000	17,370

This thesis presents a continuation of the analysis of *sorted* FF-CSB in Sanchez et al. (2005). The factors are sorted after the first phase of experimentation in order to assess the full potential of the fractional factorial step and test the validity of FF-CSB under

more realistic conditions. The assumption that the signs and/or the magnitudes of the factors can be reliably grouped or sorted prior to screening is no longer necessary. This takes a considerable burden off of the modeler or subject matter expert, especially when a large model could require categorizing of thousands of variables. If the results are useful, such a screening procedure could dramatically reduce resources required by a decision-maker.

III. EVALUATION OF *SORTED* FF-CSB

A. DESCRIPTION OF EVALUATION PROCEDURE

Initially, in order to build a clear comparison of their relative performance, *sorted* (*s*)FF-CSB is tested in the same manner under the same conditions as is the *unsorted* (*u*)FF-CSB in Sanchez et al. (2005):

- Number of factors: $K = 2^m - 1$ for $m = \{3,...,9\}$
- Maximum factor effect magnitude: $B = 5$
- *Important* threshold: $\Delta_0 = 2$
- *Critical* threshold: $\Delta_1 = 4$
- $\Pr(|\beta_i| < \Delta_0)$ = Probability of Type I error $\leq \alpha = 0.05$
- $\Pr(|\beta_i| > \Delta_1)$ = Power of detection $\geq \gamma = 0.95$
- Initial sample size: $n_0 = 5$

The stochastic main-effects model is assumed, and errors are centered on 0. For the first set of experiments the errors have constant variance $\sigma^2 = 1$. For each combination of parameters above, the factor effects are evenly distributed between 5 and -5 according to an assignment rule:

$$\beta_i = \begin{cases} -\left(-B + 2B\left(\dfrac{i-1}{K-1}\right)\right) & \text{if } i < p \\ -B + 2B\left(\dfrac{i-1}{K-1}\right) & \text{otherwise} \end{cases} \tag{5}$$

where $p = (K+1)/2$ initially. In this configuration, half the factor effects are negative. In a procedure such as CSB where no previous estimation is done either by computer or by a subject matter expert, this configuration is likely to lead to most factors, regardless of importance, being classified as *unimportant* due to cross-cancellation. Conversely, the best case scenario for CSB is when all of the factor effects have the same sign. In the

equation above, this corresponds to $p = 0$. In fact any integer $p \le (K+1)/2$ represents a configuration of factor effects where between zero and half of the effects are negative. Following Sanchez, et al. (2005), *u*FF-CSB is tested at varying values of p to simulate varying potentials for cross cancellation.

- *none negative:* $p = (K+1)/2$
- *small negative:* $p = 3(K+1)/8$
- *medium negative:* $p = (K+1)/4$
- *large negative:* $p = (K+1)/8$
- *half negative:* $p = (K+1)/2$

Figure 2 is a graphical representation of the resulting density of factor effects for $K = 63$. By design, the evaluation software switches the sign of the largest negative factor effects to positive to create the desired proportion.

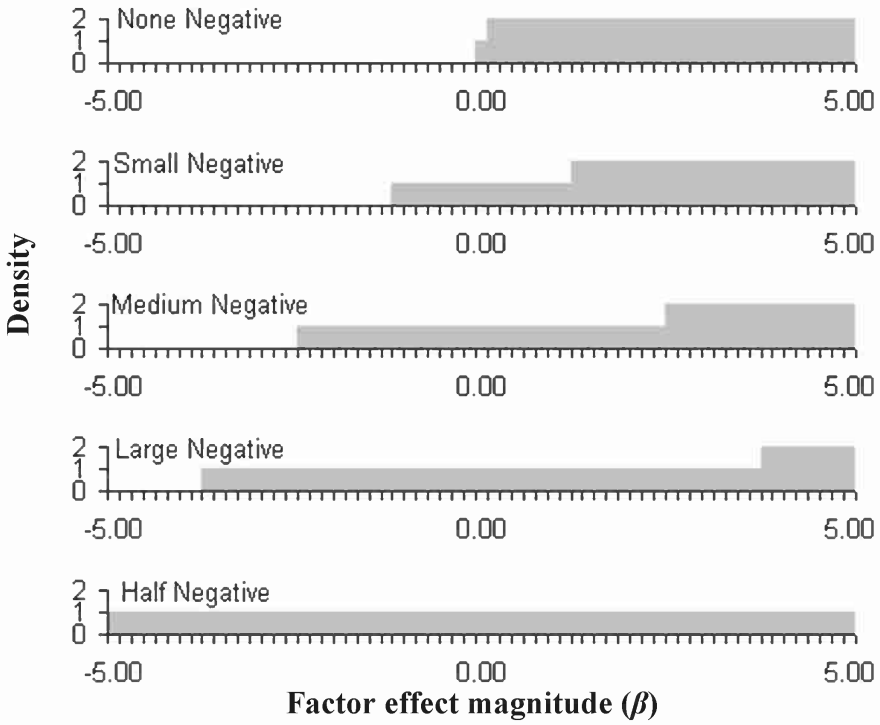

Figure 2. Density of Factor Effects; $K = 63$

This emulates the reality that estimable factor effect signs will likely belong to those factors that have the largest and most intuitive effects on the outcome. Therefore, factors with large effects are more likely to be categorized as positive or negative before CSB. For example, in a the set of factors whose magnitudes are $\{-5,-3,-1,1,3,5\}$, half of the six factors have negative magnitudes. When a fractional factorial experiment is run to classify the factors into negative or positive categories, the factors with absolute value of 3 or 5 (farther from zero) are more likely to be correctly classified, depending on the power of the test. If an alternative proportion of negative factors is desired, such as 1/3, the evaluation application will change the sign of the factor whose effect is -5 to 5, resulting in factor effects $\{-3,-1,1,3,5,5\}$. By this convention, the relative sizes of the estimated positive/negative groups are more likely to represent the desired proportion. Alternatively, if the sign of the factor with the smallest effect is switched, the probability is higher that the resulting positive/negative groups will be identical to the same groups when no signs are switched and half of the effects are negative.

Proportions of factor effects by magnitude are roughly constant and depend on the threshold values relative to the maximum factor effect magnitude. For the configurations in these experiments, c*ritical* factors make up approximately 20% of total factors, while *important* and *unimportant* factors make up approximately 40% each. Since there are necessarily an integer amount of factors in each category, these proportions are more accurate for larger values of *K*.

As in Sanchez et al. (2005), the *s*FF-CSB procedure is run 400 times at each combination of parameters where $K = \{255, 511\}$ and 1,000 times each for smaller values of *K*. For each factor and screening procedure run, a binary output indicates whether each the factor was classified as *important.* These binary outputs can be aggregated to a binomial probability of detection over all factors in a specific category (approximately $0.2K$ critical factors and $0.4K$ important or unimportant factors). The number of replications guarantees that the standard deviations for replication probabilities are small, less than $\sqrt{0.5^2/(0.2K \cdot 400)} \le .004$ for $K = \{255, 511\}$ and $\sqrt{0.5^2/(0.2K \cdot 1000)} \le .013$ for the smaller values of *K*. In this paper, the experimental standard deviations are

considered to be negligible when compared to the general results. Accordingly, only the mean detection rates are reported.

Intuitively, based on the design and success of the fractional factorial experiment, the first phase of *s*FF-CSB should provide two groups of factors in which relatively unimportant factors are concentrated at one end of the group. One group consists of factors whose effects are estimated to be negative; the other, factors whose effects are estimated to be positive. In the second phase, this arrangement should lead to identification of *critical, important,* and *unimportant* factors that is at least as accurate as *u*FF-CSB, but with fewer runs required. A positive result would indicate that given the underlying assumptions and in the absence of previous knowledge of the signs or magnitudes of factor effects, *s*FF-CSB is a robust and efficient screening tool for simulation experiments.

The results of this evaluation are summarized in Table 3. For each screening procedure and parameter combination, the proportion of *critical, important,* and *unimportant* factors correctly classified is shown (1.000 is ideal). In this analysis, a correct classification for *important* and *critical* factors means classification as *important.* The average runs required is also displayed for each parameter combination.

B. PRELIMINARY RESULTS

To summarize prior research ('CSB' column of Table 2), Sanchez et al. (2005) find that the performance of CSB is highly dependent on the proportion of negative effects in the experiment. Although the procedure exceeds power and error specifications when none or a small number of effects are negative, the correct classification rates across all categories deteriorate as the number of negative factors approaches the number of non-negative factors. When half of the factors are negative, the procedure is unable to produce useful results. The number of runs required for CSB also decreases as more negative factors are introduced because more factors are discarded, often incorrectly, due to cancellation. In general, while CSB performs well when the conditions are optimal, it cannot classify factors reliably in the general sense, and any gains in run requirements as negative factors increase are at the cost of accuracy.

In contrast, due to separation by estimated sign after the first stage, the *u*FF-CSB procedure is not as susceptible to cancellation between positive and negative factors during estimation. The procedure exceeds power and error specifications in all cases, regardless of the number of negative effects. *Critical* and *unimportant* effects are correctly classified at least 99.7% of the time, and correct classification rates for *important* factors range from 42.7% to 80.9%. However, the number of runs required for *u*FF-CSB is generally larger. This is generally true for one primary reason: since the *u*FF-CSB has grouped factors by their estimated signs, far fewer factors are discarded as *unimportant* due to cancellation. Consequently, fewer factors are classified per bifurcation step and more steps are required. Overall, *u*FF-CSB clearly outperforms CSB. It requires very few additional runs even when the CSB assumption of no negative factor effects is satisfied, and efficiently and successfully classifies factors in situations where CSB breaks down.

C. EFFICIENCY OF *SORTED* FF-CSB

The results of the new experiments (conducted with the MATLAB® code used by Sanchez et al. 2005) appear in Table 3. The performance of *u*FF-CSB is also included to facilitate comparisons. Like *u*FF-CSB, the *s*FF-CSB procedure exceeds power and error requirements in all cases, regardless of the number of negative factors. In all cases, correct classification rates for *critical* factors are the same as or slightly lower than those of *u*FF-CSB. For *unimportant* factors, all but seven cases show no improvement or a slight decrease in correct classifications. However, the minimum correct classification rates among *critical* and *unimportant* factors are still high at 99.4%; these are substantially higher than the user-specified requirements of 95%. Classification rates for *important* factors are also slightly lower than those of *u*FF-CSB in all but three cases and range from 42.4% to 80.8%. Overall, correct classification rates by category for the two FF-CSB procedures are statistically indistinguishable.

Table 3. Performance of *u*FF-CSB and *s*FF-CSB

Pattern of β values	K	*Unsorted* FF-CSB Correct Class. Prop.			Avg.	*Sorted* FF-CSB Correct Class. Prop.			Avg.	Relative Efficiency of *s* FF-CSB
		Crit.	Imp.	Unimp.	Runs	Crit.	Imp.	Unimp.	Runs	
None Negative	7	1.000	0.809	0.999	110	1.000	0.808	0.999	108	0.98
	15	1.000	0.432	1.000	268	0.998	0.428	1.000	219	0.82
	31	0.998	0.493	0.999	656	0.997	0.490	1.000	528	0.81
	63	0.999	0.490	1.000	1,563	0.999	0.481	1.000	1,208	0.77
	127	1.000	0.507	1.000	3,692	0.999	0.502	1.000	2,817	0.76
	255	1.000	0.497	1.000	8,704	1.000	0.487	1.000	6,484	0.74
	511	1.000	0.497	1.000	19,528	1.000	0.492	1.000	14,584	0.75
Small Negative	7	1.000	0.809	0.999	110	0.999	0.804	0.999	107	0.98
	15	0.999	0.431	1.000	251	0.996	0.419	1.000	226	0.90
	31	0.999	0.498	0.999	605	0.997	0.485	1.000	509	0.84
	63	1.000	0.487	1.000	1,461	0.999	0.481	1.000	1,217	0.83
	127	1.000	0.506	1.000	3,421	0.999	0.498	1.000	2,708	0.79
	255	1.000	0.496	1.000	8,024	1.000	0.489	1.000	6,268	0.78
	511	1.000	0.499	1.000	18,132	1.000	0.493	1.000	14,205	0.78
Medium Negative	7	1.000	0.804	0.999	100	1.000	0.806	0.999	97	0.97
	15	0.999	0.432	1.000	250	0.996	0.424	0.999	209	0.84
	31	0.999	0.502	0.999	586	0.998	0.490	0.999	491	0.84
	63	0.999	0.494	1.000	1,407	0.998	0.487	1.000	1,084	0.77
	127	1.000	0.508	1.000	3,307	0.999	0.503	1.000	2,512	0.76
	255	1.000	0.496	1.000	7,550	1.000	0.492	1.000	5,767	0.76
	511	1.000	0.499	1.000	17,703	1.000	0.495	1.000	13,083	0.74
Large Negative	7	1.000	0.785	0.999	100	1.000	0.800	0.998	98	0.98
	15	0.999	0.427	1.000	234	0.996	0.425	1.000	199	0.85
	31	0.997	0.495	0.999	558	0.996	0.490	1.000	492	0.88
	63	0.999	0.489	1.000	1,329	0.998	0.476	1.000	1,093	0.82
	127	1.000	0.505	1.000	3,171	0.999	0.497	1.000	2,549	0.80
	255	1.000	0.494	1.000	7,440	1.000	0.486	1.000	5,797	0.78
	511	1.000	0.499	1.000	17,595	1.000	0.491	1.000	13,314	0.76
Half Negative	7	1.000	0.770	1.000	99	1.000	0.788	1.000	88	0.89
	15	0.999	0.431	1.000	231	0.997	0.424	1.000	226	0.98
	31	0.998	0.494	0.999	556	0.994	0.486	1.000	454	0.82
	63	0.999	0.489	1.000	1,313	0.997	0.479	1.000	1,085	0.83
	127	1.000	0.507	1.000	3,143	0.999	0.497	1.000	2,410	0.77
	255	1.000	0.495	1.000	7,441	0.999	0.486	1.000	5,656	0.76
	511	1.000	0.499	1.000	17,370	1.000	0.491	1.000	13,095	0.75

In contrast, experimental run requirements for *s*FF-CSB are significantly less in all cases. *s*FF-CSB achieves 9.13% to 21.97% improvement for small K values ($K = \{7,15\}$) and 17.61% to 28.99% improvement for larger values of K. Relative improvement in runs over *u*FF-CSB generally increases as K increases, indicating that the advantage of *s*FF-CSB is actually more pronounced as the simulation model becomes larger and/or more complex.

As a representation of the general behavior of run times, Figure 3 shows the distributions of runs required for all proportions of negative factors when $K = 63$. The distributions of run times in the second phase appears to be bell-shaped, although skewed to the right. In each of the samples below, the median is below the mean. More complete analysis of the effects of negative factors on run times is discussed later.

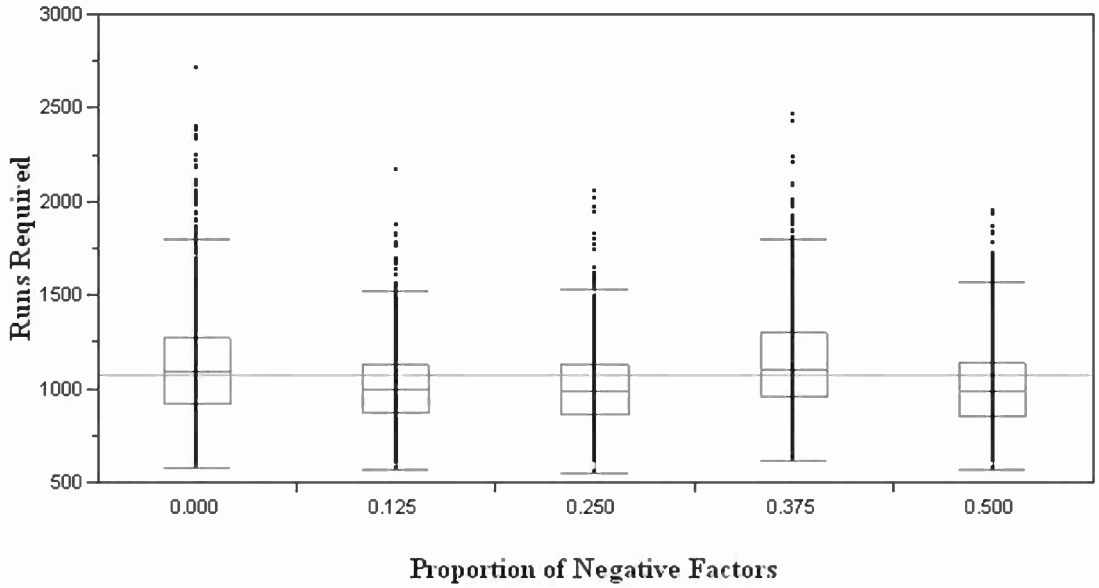

Figure 3. Experimental Runs Required by Proportion of Negative Factors, $K = 63$

D. PROBABILITY OF DETECTION VS. FACTOR EFFECT MAGNITUDE

The next round of analysis is an investigation of classification rates as a function of factor effect magnitude (β_i) under varying user- and model-determined conditions. Sanchez et al. (2005) provide this functionality in their evaluation applications. Each experimental replication outputs a binary figure indicating whether or not each factor was identified as *important*. Assuming a binomial distribution, these can be aggregated over many replications into a probability of classification as *important* for a given set of parameters. Intuitively, factors with larger effect magnitudes will be more likely to be classified as *important.* This is supported by the performance results above, where both *critical* factors (large β_i) and *unimportant* factors (small β_i) are almost always identified correctly. The actual shape of this function, however, is also of interest. Figure 4 shows the observed probability of classification as *important* as a function of $|\beta_i|$ when $K = 63$, half of the factor effects are negative, and factor effects are distributed evenly on the interval $[-B, B] = [-5, 5]$.

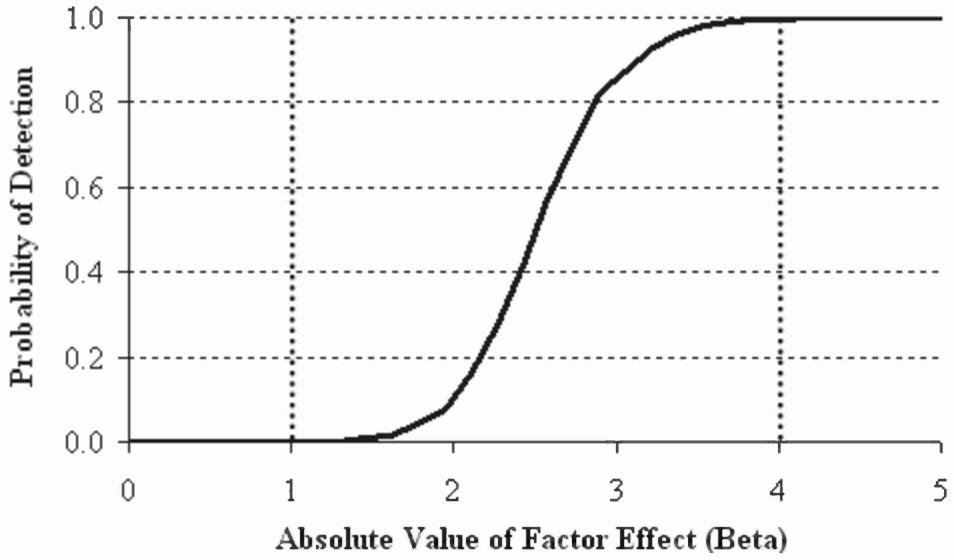

Figure 4. Probability of Detection vs. Factor Effect Magnitude (Beta) K= 63; Proportion of Negative Factors: Half; $\Delta_0 = 1$; $n = 1000$

For factor effects in the *unimportant* category ($|\beta_i| < \Delta_0$), the observed probability of classification as *important* is effectively 0, the best possible. Similarly, for effect

magnitudes in the *critical* range ($|\beta_i| > \Delta_1$), the observed classification rate in this range is 1, also the best possible. For effect magnitudes between the two threshold values, the probability curve is s-shaped, although the underlying behavior of the curve is not discussed in this analysis.

E. CHANGING *IMPORTANT* AND *CRITICAL* THRESHOLDS

A related experiment analyzes *s*FF-CSB performance when the lower threshold (Δ_0) is varied and all other parameters are held constant. The experiment above is rerun with $\Delta_0 = 2$ and $\Delta_0 = 3$. Figure 5 shows the observed probability of classification as *important* as a function of β_i for the three threshold values. Classification probabilities for all three thresholds follow the same pattern of essentially perfect for *critical* and *unimportant* factors with a nonlinear s-shaped function in between. Increasing Δ_0 lowers the curve, showing that the classification is more difficult when the difference between the important and critical thresholds shrinks.

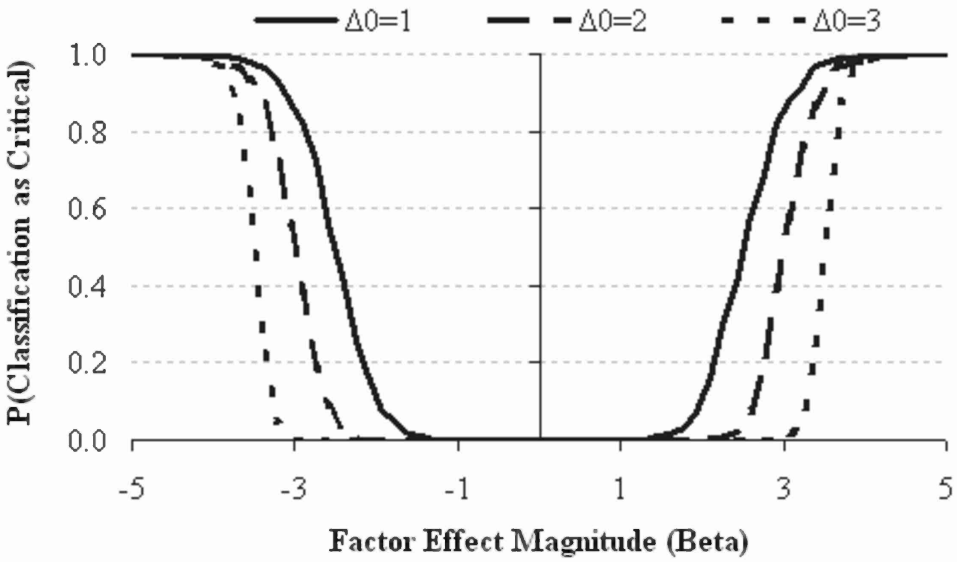

Figure 5. Probability of Factor Classification as Important vs. Beta, *K*=63, Equally Spaced Factor Effect Magnitudes

The effects of sparsity on probability of detection are also explored by running the same three experiments above with sparse effects. To simulate sparsity, all $\beta_i < \Delta_1$ are

set equal to 0, and all $\beta_i \geq \Delta_1$ are set to $B = 5$. Under these conditions, classification rates are nearly perfect: *s*FF-CSB fails to identify only 2 of 12,600 individual *critical* factors when $\Delta_0 = 1$ and half of the factor effects are negative. The number of runs required to complete *s*FF-CSB is also analyzed for multiple factor arrangements. Since at each bifurcation step the CSB procedure takes individual samples from a simulation model sequentially (after an initial sample size, n_0) until power and error specifications are met, the number of model runs required is highly dependent on the characteristics of the group being sampled. A group consisting of one factor with a small effect variance will take fewer runs to classify than a group consisting of many factors with disparate means and large variance. Over an entire FF-CSB procedure consisting of many hundreds or thousands of CSB replications, these differences should be substantial and may be predictable for certain parameter levels. The detection experiments above represent variations in Δ_0 (three levels) and sparsity (two levels). For the following experiments, the proportion of negative factors is also varied (two levels), for a total of twelve experiments representing 12,000 data points. For each experiment, Figure 6 shows a 95% confidence interval on the mean number of runs. Regression shows that while the number of negative factors does not affect the number of runs ($p = .56$), both sparsity and Δ_0 have a large significant effect, as does their interaction ($p \approx 0$ in all cases).

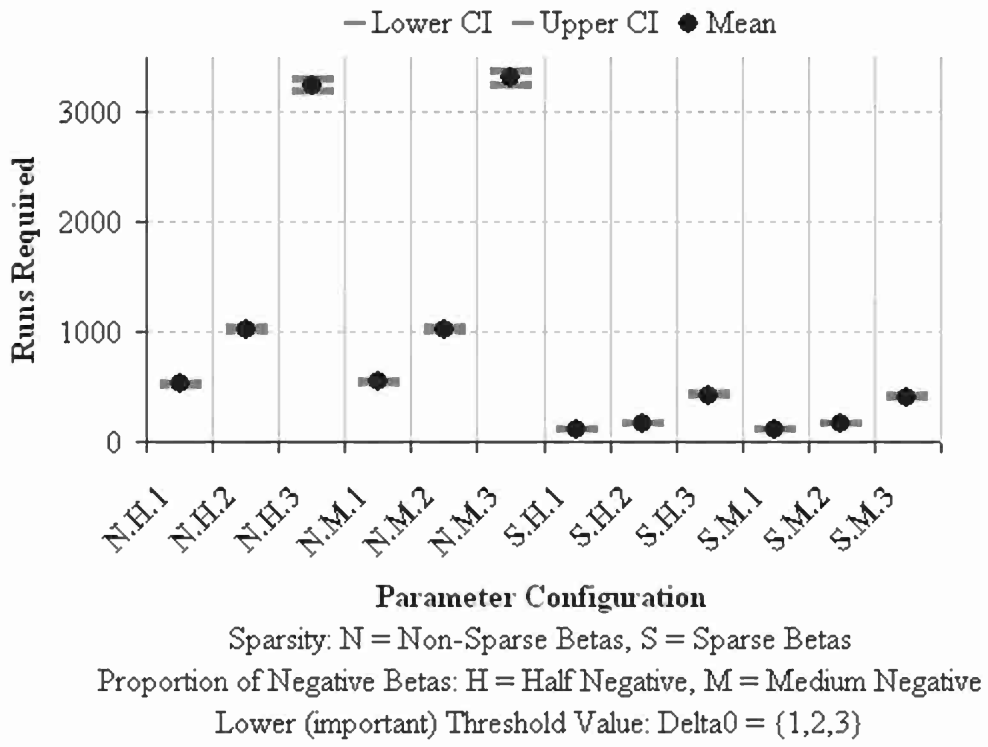

Figure 6. Confidence Intervals for Mean Runs Required by *s*FF-CSB

F. INCREASING RESPONSE VARIANCE

A further test of the capability of *s*FF-CSB is increased response variance. While the variances are still constant as in previous experiments, the procedure is run for $\sigma^2 = \{1^2, 5^2, 10^2\}$ and $\Delta_0 = 1$. Since CSB uses hypothesis testing to classify group effects where μ and σ are unknown, the width of the confidence interval is proportional to $s/\sqrt{n}$, where s is approximately σ. Consequently, for constant variance σ^2, the mean number of runs required during the final bifurcation step should be approximately σ^2 times this number of runs when $\sigma^2 = 1$, although in practice the *overall* impact of increasing or decreasing the variance is not as straightforward to compute. In this experiment, for example, a predetermined minimum number of samples taken at each CSB replication ($n_0 = 5$) may cause unnecessary sampling for groups of factors with small response variance, causing the number of runs required to be unnecessarily large. Alternatively, a high variance case may require more than n_0 samples at each step,

resulting in a runs required figure which is more accurate. Figure 7 shows confidence intervals on the mean number of runs required for $\sigma^2 = \{1^2, 5^2, 10^2\}$. The mean number of experimental runs required when $\sigma^2 = 25$ and 100 is roughly 22 and 89 times the mean number of runs required when $\sigma^2 = 1$, respectively.

This small experiment illustrates that while larger variance has an extreme effect on required runs, the effect may be hard to predict based on variance alone. Other considerations, such as the patterns of the factor effects and other procedure parameters, likely play significant roles.

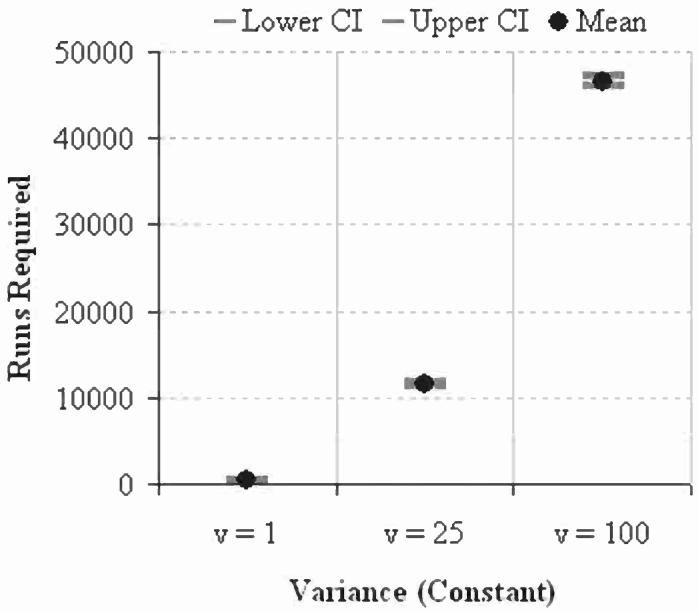

Figure 7. Confidence Intervals for Mean Runs Required by *sorted* FF-CSB

G. HETEROGENOUS RESPONSE VARIANCE

As stated above, evaluation of FF-CSB up to this point has assumed that the variance of the model response is constant regardless of factor magnitudes or grouping. While this assumption is a major simplification of most useful simulation models, it is valuable in isolating the performance effects of the parameters tested above. Introduction of response variance that is other than constant will require detailed study based on a wide range of assumptions. A major consideration for these experiments is the nature of the relationship, if any, between the magnitude of factor effects and their variance. In

practice, the distribution of a factor effect in reality or in simulation will depend on its function in the model as well as the levels of other factors in the current simulation run. Response variances can usually be reasonably bounded or parameterized based on the scenario being modeled. The distribution of a physical response such as projectile time of flight may have very small variance, whereas a qualitative response such as enemy aggression level may be nearly impossible to predict and be widely distributed. A simple assumption suggested by Sanchez et al. (2005) is that the variance of a stochastic model response increases as the mean response increases. This is a common occurrence in practice, but it represents a very difficult scenario for the screening procedure. If the response standard deviation is proportional to the response mean, then all design points will have identical *p* values when differentiating the sample response from 0.

To quantify the effect of heterogeneous errors on *s*FF-CSB performance, the procedure is run against the model:

$$Y = f(x_i) + \varepsilon, \text{ where } \varepsilon \sim N\left(0, \left(m \cdot f(x_i)\right)^2\right) \text{ and } f(x_i) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i . \quad (6)$$

For each sample, the model response is sampled from a normal distribution centered on a linear combination of the factor levels and their effects. The response standard deviation is $m \cdot f(x_i)$, where *m* is an arbitrary multiplier. For example, if the expected response is 20 and $m = .50$, the experimental factor effect will be normally distributed about 20 with $\sigma = .5 \cdot 20 = 10$. The experiment is designed to extract the basic relationship between σ and run requirements and to compare constant and proportional σ values that, for these parameters, require the same average number of runs. The smallest value possible for *m* is 0, and a small initial experiment provides a maximum value for *m* (*m* = .20) that produces average run requirements slightly larger than those required by the largest constant variance *v* = 100. This provides a common range of run requirements with which to compare the two scenarios. The procedure is run once for each multiple of 0.0001 from 0.0000 to 0.2000 (*n* = 2001). The resulting run times and polynomial fit are displayed in Figure 8. The heavy dotted lines represent the mean runs required when σ^2 is constant at 25 and 100.

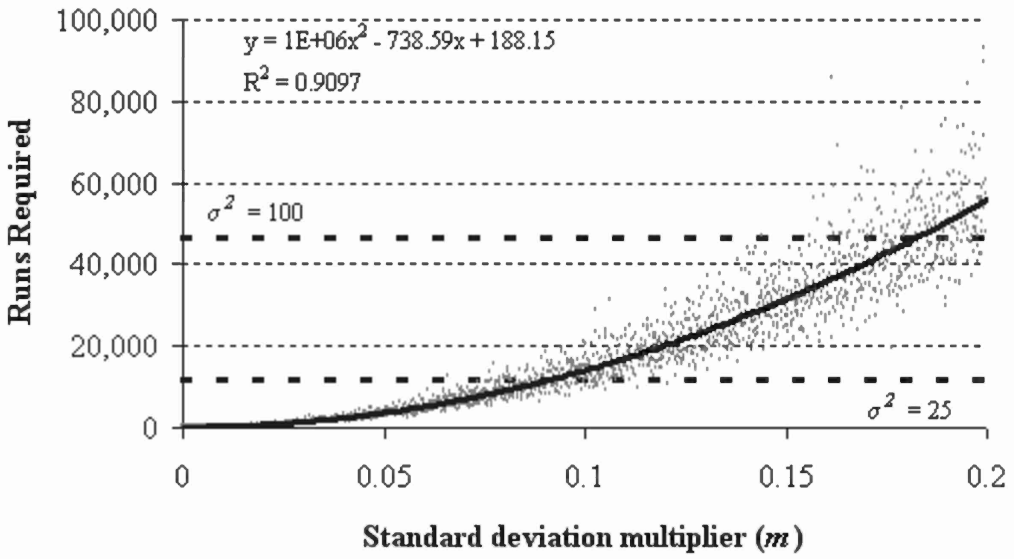

Figure 8. Runs Required vs. σ Multiplier (Proportional σ)

A comparison of run times for cases of constant and proportional σ is desired. While in the constant case, a fixed σ provides a baseline figure for comparison, a similar figure for the proportional case has to be computed. To accomplish this, the original set of factor effect magnitudes is randomly reordered several times. Each permutation of factor effects is multiplied by the design matrix of the original fractional factorial experiment to extract a large sample of model responses, providing a mean response. For each experimental value of *m*, therefore, an equivalent constant variance scenario can be identified and the run times compared. For example, the mean response for this set of factors is $\bar{Y} \approx 23.3$. Using this $\bar{Y}$, the bottom axis in Figure 5 can be transformed to represent the equivalent σ for this case. The resulting plot is shown as Figure 9.

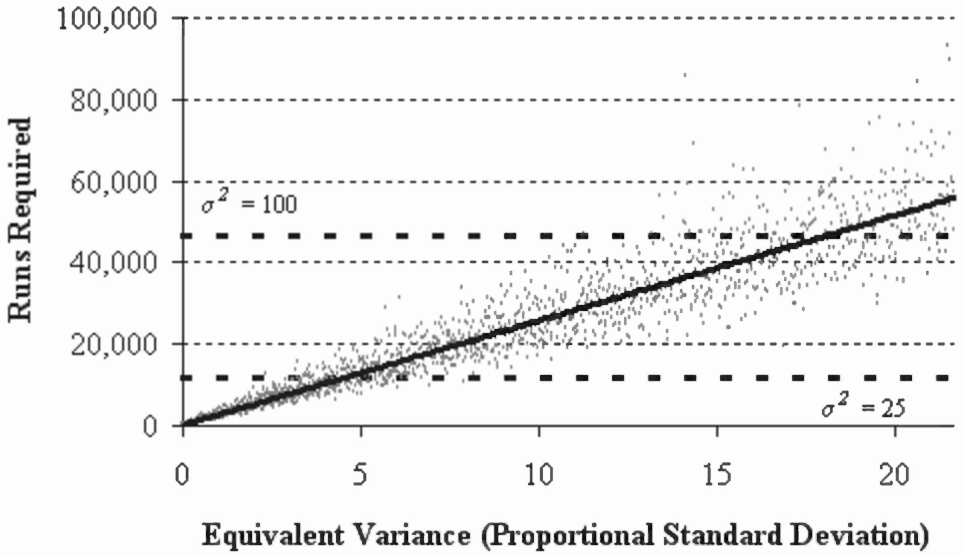

Figure 9. Runs Required vs. Equivalent Variance (Proportional σ)

The plot suggests that for an equivalent variance, more runs are required to classify factors if the response standard deviation is proportional to the response. For example, when σ is proportional to Y and the average response variance is 5 (m = .215), the average number of runs required is similar to a constant variance case where response variance is fixed at 25. Although proportional σ in responses may represent an extreme case, a strong increase in run requirements will also be apparent in cases of less challenging response distributions.

Additionally, these analyses show that the shape of the detection probability curve has stayed relatively stable across multiple scenarios. For example, Figure 10 shows the probability of detection curve for proportional σ when m = .20. The only consequence in prediction rates themselves appears to be that prediction rates are more volatile for factors with larger effect magnitudes, since these will consistently belong to groups whose overall response is high and therefore widely distributed.

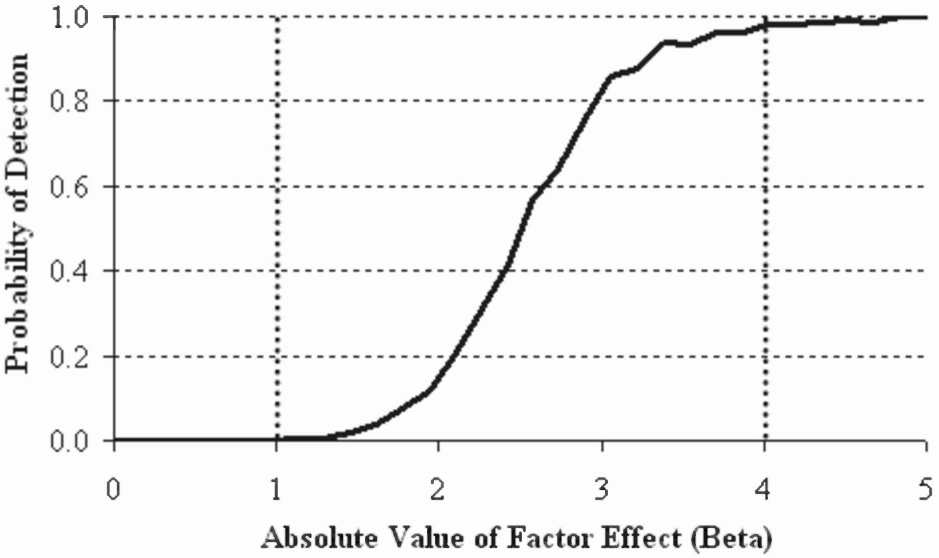

Figure 10. Probability of Detection vs. Factor Effect Magnitude (Beta). *K*= 63; Proportion of Negative Factors: Half; $\Delta_0 = 1$; m = .20; n = 500

IV. DISCUSSION

With an analyst as an end user, a screening procedure that finds widespread application should come with a comprehensive understanding of strengths, weaknesses, and limitations. For example, given that trustworthy results are expected, an analyst should be able to have a good idea of how long a screening procedure will take. In some situations, especially when analysis time is limited, it may not be feasible to run a screening experiment involving *all* potential factors—even though screening experiments can be extremely efficient and provide better insights to the decision maker. An analyst may also forgo using a particular screening procedure if he has reason to believe the model being screened exploits a weakness in the procedure, such as an inability to handle factor effects of different signs or non-constant response variance, that will cost significantly in accuracy or run time.

To this end, this analysis represents a more in-depth evaluation of *s*FF-CSB that not only shows improvement over *u*FF-CSB, but attempts to parameterize performance at least qualitatively for a variety of different settings. Much more experimentation is required before an end user can make qualified statements as to the run time of a proposed screening experiment. However, this procedure is valuable for defense analysts because it provides a systematic way of investigating simulation models of military operations—which often contain hundreds or thousands of factors. When a long list of potentially important factors can be quickly trimmed to a short list of demonstrably important factors, analysts can apply a greater portion of their time, effort, and computing resources toward higher-resolution experiments that focus on the factors that matter.

Based on preliminary results shown in Table 2, *s*FF-CSB is a viable alternative to *u*FF-CSB in cases where the signs or magnitudes of factor effects cannot be reliably pre-determined. As expected, run times are significantly smaller than those of *u*FF-CSB with no sacrifice in detection rates for important factors. While the desire for shorter experiment preparation and run times requires no discussion, the removal of pre-determination of factor effects introduces a further consideration. Where factor effects are pre-determined in sign, a screening procedure may only be looked to for magnitude estimation or classification by category as in CSB. An experiment such as *s*FF-CSB,

however, relaxes the need for prior knowledge regarding the signs of potential factor effects and moves the utility of the procedure away from validation. A procedure that requires only a model as input is closer to screening in the purest sense.

As shown in Figure 1, the probability of detection for factors is a function of the effect magnitude relative to the thresholds Δ_0 and Δ_1 set by the user. Under the assumptions of this analysis, the shape of the function is stable as thresholds are varied, although introduction of higher response variance appears to cause higher variance in the mean prediction rate for a given factor effect (Figure 6). Further analysis of this function would incorporate more challenging arrangements of β_i values, perhaps where effect magnitudes are concentrated around threshold values. Most importantly, though, closer threshold values are associated with larger run times. When run times are parameterized only by the value of the lower threshold, the function is strongly quadratic, although further analysis will likely find that this relationship is dependent on multiple factors, including response variance and the arrangement of β_i values. The significance to the analyst is that the size of the indifference zone matters, and that there is a trade-off between shorter run times and consistent removal of unimportant factors. Although it may save runs in further experiments, reliably isolating fewer factors as important in *s*FF-CSB will cost significantly in run times.

The more general run time experiment summarized by Figure 3 explores the effect on run times when the number of negative factors is changed and sparsity is introduced. The results are significant and intuitive in the case of sparsity, but may be incomplete where negative factors are concerned. Furthermore, unlike threshold values, these parameters do not represent user inputs, but rather characteristics of the model. Where the screening procedure is used without little or no prior knowledge about the system behavior, the analyst is less likely to consider these factors when considering a screening procedure.

Consideration of the response distribution should be important to the analyst when preparing a screening simulation. Even when the expected variance of the response is constant (or thought to be constant), run times are significantly effected by variations simply because more samples are needed to meet power and error specifications. As stated above, the mean number of runs required depends on several factors but increases

with the response standard deviation. When considering a screening procedure, the analyst will likely be able to extract an idea of response variance from the model code itself (especially when error is explicitly parameterized in the model), or can generate at least a few responses with which to estimate a response distribution.

In cases where the response variance is not thought to be constant, getting some idea of expected run times is currently much harder. While, at this point, non-constant variance appears to be detrimental to run times, especially when higher variances are associated with higher responses, more comprehensive experimentation is required. Not only are there many possible general relationships between mean response and response variance, an analyst may also find cases where estimating such a relationship is impossible. The proportional σ used in this analysis represents an extreme case, and evaluating the effects of response/variance relationships may be most helpful to analysts if it is focused separately on relationships most often encountered in practical simulation applications.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Bettonvil, B. and J. P. C. Kleijnen. 1997. Searching for important factors in simulation models with many factors: sequential bifurcation. *European Journal of Operational Research* 96: 180-194.

Cheng, R. C. H. 1997. Searching for important factors: sequential bifurcation under uncertainty. In *Proceedings of the 1997 Winter Simulation Conference*, eds. S. Andradottir, K. J. Healy, D. H. Withers, and B. L. Nelson, 275–280. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available via <http://www.informs-sim.org/wsc97papers/0275.PDF> Last accessed June 2006.

Kleijnen, J. P. C., B. Bettonvil, and F. Persson. 2005. Finding the important factors in large discrete-event simulation: sequential bifurcation and its applications. In *Screening*, eds. A. M. Dean and S. M. Lewis. Screening. New York: Springer-Verlag.

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. 2005. A user's guide to the brave new world of simulation experiments. *INFORMS Journal on Computing* 17: 263–289.

Lucas, T. W., S. M. Sanchez, L. Brown, and W. Vinyard. 2002. Better designs for high-dimensional explorations of simulations. In *Maneuver Warfare Science 2002*, eds. G. Horne and S. Johnson, 17-46. Quantico, Virginia: USMC Project Albert.

MATLAB® 7.0.1, 2004. Natick, Virginia: The MathWorks, Inc.

Sanchez, S. M., T. W. Lucas, and H. Wan. 2005. A two-phase screening procedure for simulation experiments. In *Proceedings of the 2005 Winter Simulation Conference*, eds. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 223-230. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available via <http://www.informs-sim.org/wsc05papers/023.pdf> Last accessed June 2006.

Wan, H., B. Ankenman, and B. L. Nelson. 2003. Controlled sequential bifurcation: a new factor-screening method for discrete-event simulation. In *Proceedings of the 2003 Winter Simulation Conference*, eds. S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, 565-573. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available via <http://www.informs-sim.org/wsc03papers/070.pdf> Last accessed June 2006.

Wan, H., B. Ankenman, and B. L. Nelson. 2006. Controlled sequential bifurcation: a new factor-screening method for discrete-event simulation. *Operations Research*, forthcoming.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Professor Susan Sanchez
   Naval Postgraduate School
   Monterey, California

4. Professor Hong Wan
   Purdue University
   West Lafayette, Indiana